Experience with Real Time Event Reconstruction Farm for Belle Experiment

R.Itoh, M.Nakao, T.Higuchi, S.Y.Suzuki, I.Adachi and N.Katayama
KEK, Tsukuba, Japan

Abstract

The luminosity of the KEKB accelerator is increasing very rapidly and the consumption of the data recording media of the Belle experiment may exceed a manageable limit in the near future. The data used in the offline analysis are only 25% of taken data and the reduction of the data to this level before recording is desired to keep the storage size within the manageable limit. However, a massive computing power is required to perform the offline-level event selection. For this purpose, a real time event reconstruction farm (RFARM) has been built and implemented in the Belle DAQ system. The farm consists of 85 processing nodes each of which is equipped with dual Athron processors running at 1.6GHz, and a set of control nodes. RFARM is working stably in the current DAQ condition at an average L1 trigger rate of 450Hz with an event size of 40kBytes. The real time data monitoring using RFARM is introduced based on the asynchronous transfer of histogram contents placed on a shared memory over the network. It enables the detailed check of the data quality during the data taking.

INTRODUCTION

The Belle experiment, which is a B-factory experiment at KEK in Japan, is now acquiring more and more data for the further study of the CP violation in B meson decays. The KEKB accelerator has achieved the world's highest luminosity of $1.39 \times 10^{34}\,\mathrm{cm}^{\Gamma 2}\mathrm{sec}^{\Gamma 1}$ and the luminosity is still increasing.

Fig. 1 shows the data acquisition system of the Belle experiment. The system is based on the Q-to-T conversion with the FASTBUS TDC readout (except for the readout of the silicon vertex detector) and the switchless event building farm[1]. Since the readout electronics is not pipelined, the system has an intrinsic readout deadtime of $\sim$3The level 2.5 trigger is performed on the event building farm whose reduction factor is about 40%.

Recently, the KEKB accelerator is operated in the "continuous injection" mode. In this operation, the electron and positron beams are continuously injected alternatively to the accelerator rings while the data taking is going on. Fig. 2 shows the accelerator operation with and without the continuous injection. The left figure shows the operation without the continuous injection. The beam current and luminosity go down during the data taking in this case. The right figure is with the continuous injection. Since the beam currents are kept almost constant, the luminosity does not

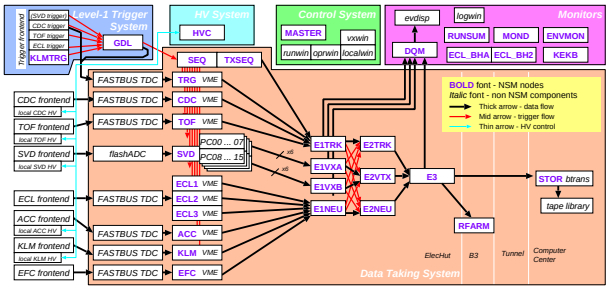

Figure 1: The Belle Data Acquisition System.

go down. Although the continuous injection requires an additional DAQ deadtime for the veto during the actual beam injection (3.5msec per each injection), the integrated luminosity gains more than 30%.

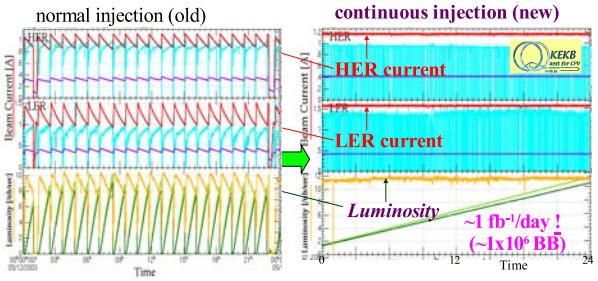

Figure 2: The accelerator operation with (right) and without (left) the continuous injection. The lines in red show the beam currents and the lines in yellow show the luminosity.

As a consequence of the luminosity increase, the consumption of data recording media has been multiplying from the beginning of the experiment and may exceed a manageable limit in the near future. The data used for the physics analysis are selected in the recorded data after the full event reconstruction. The fraction of selected events is about 25% of the recorded data. If the selection can be done before recording, we can reduce the mass storage consumption drastically. For this purpose, the Real Time Event Reconstruction Farm (RFARM) has been developed and added to the Belle data acquisition system. The detail of the architecture and operation of RFARM is reported in this talk.

ARCHITECTURE

RFARM consists of an input distributor node, an output collector node with a fast disk system, a set of control nodes

(Figure 3), and a large number of event processing nodes. All the nodes are PC servers with dual CPUs and operated under Linux (RedHat7.3). The data flow of RFARM is managed by a set of small programs, namely, **receiver**s, **transmitter**s, a **writer** and **ring-buffer**s (Fig. 4). The **ring-buffer**s are widely used to absorb a sudden change in the data taking condition for a short period.

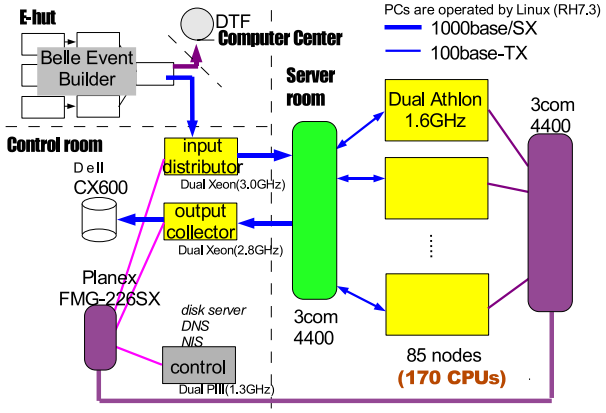

Figure 3: The hardware configuration of RFARM.

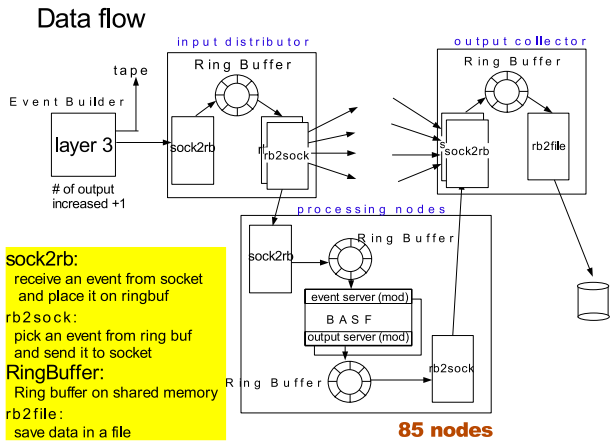

Figure 4: The data flow in RFARM.

The data from the event building farm are sent to the input distributor node via a point-to-point socket connection over a 1000base-SX optical fiber. The data are read by a **receiver** and placed in a **ring-buffer** on a shared memory. A set of **transmitter**s then pick up events from the **ring-buffer** and send them to corresponding event processing nodes via socket connections. The input distributor and event processing nodes are both connected to a big network switch via 1000base-SX and 100base-TX, respectively, and events are distributed to processing nodes through the switch. Currently 85 processing nodes are connected to the switch.

An event processing node is equipped with two Athron processors operated at 1.6GHz and the full event reconstruction is performed utilizing both processors. Event data are read from a receiving socket by a **receiver** and placed on a **ring-buffer**. The events are then read by the Belle AnalysiS Framework(BASF)[2] with a ring-buffer interface. The same reconstruction program as that used in offline is executed in BASF. The processing time for an event is widely changed according to the event type. The typical time is a few seconds per processor. Processed data are then written to a different **ring-buffer**. A **transmitter** picks up one event and sends it back to the network switch by sharing the connection with the input by assigning a different IP address on the same NIC. The data are then sent to the output collector node through a different 1000base-SX connection from the network switch.

A set of **receiver**s running on the output collector node corresponding to each of event processing nodes receive the processed data and place them in a **ring buffer**. The data from all processing nodes are collected on the buffer and written to a fast disk by a **writer**. The disk is an array of fiber channel disks (Dell CX-600) whose writing speed is more than 100MB/sec.

The control of the RFARM nodes is done using the NSM (Network Shared Memory)[3] package which is a home-grown slow-control software capable of passing messages and sharing data over a control network. The network is separated from the one used for the data flow. The operation of RFARM nodes is made by exchanging messages with a control node. The histogram accumulated on event processing nodes are periodically collected and placed on the shared memory of the control node for the purpose of the real time data monitoring.

OPERATION

RFARM started operation with 29 processing nodes in Oct. 2003. The operation was mainly for the test of the data flow in the beam runs and the event reconstruction was not performed. The true full event reconstruction was started on from Feb. 2004 with 43 nodes. To keep up with the increase in the accelerator luminosity, 32 more nodes were added time by time and the farm is now operated with 85 processing nodes.

In a typical condition of the accelerator operation, the average L1 trigger rate is about 450Hz with an event size of about 40 kbytes. The performance of RFARM in this condition is shown in Fig. 5. The farm is operated with 71 processing nodes when the performance is measured. The event selection is not yet performed and all the events from the event building farm are processed and recorded. As seen from the figure, the average processing rate is about 230Hz (after the level 2.5 trigger) and the average recording rate is 21MB/sec.

Since the trigger rate and data size largely fluctuate during the data taking and also the processing time per event strongly depends on the event type, the averaging of the data flow rate is necessary for a stable operation. For this purpose, ring buffers with variable record length are extensively used in the data flow. In particular, the depth of input ring buffer in each processing node is set at 128Mbytes.

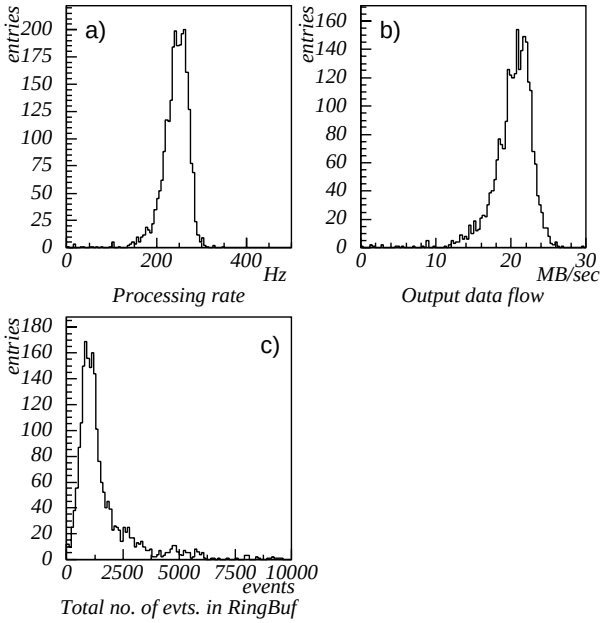

Figure 5: Performance of RFARM measured for one day operation. a) processing rate, b) output data flow, and c) total number of events in input ring buffers.

Since 71 nodes are in operation, the total size is about 9GB which is capable of storing data for 10 minutes. The distribution of the total number of events queued in the input ring buffers is shown in fig. 5-c). The typical number is about 1000 and sometimes becomes more than 5000 when the background condition becomes worse. This shows a large fluctuation in the data flow rate due to the change in the accelerator condition is well absorbed by the ring buffer scheme.

When starting a run, the begin run processing is performed in each processing node. The actual processing is started when a begin_run record generated by the event building farm is received. The record is replicated and sent to all processing nodes by the input distributor. Since the off-line DST production code is executed on the processing node without any modifications, the begin run processing includes a heavy access to the constant database. In the off-line environment, one postgreSQL server is operated and all the database accesses are centralized to this server. However, if the same scheme is used in RFARM, the database is accessed from 85 nodes at one time which makes the begin run processing significantly slow. To avoid this, the postgreSQL is operated on each processing node with replicated constant data and all database accesses are localized. The time to start a run is less than 30 seconds with all the begin run processing.

The end run processing includes the merging of log files, histogram files, and event statistics generated by each processing nodes. They are done after a run is stopped and independently of the DAQ processing for next run using a spooling scheme. This also reduces the time for stopping a run to less than 30 seconds.

REAL TIME DATA MONITORING

For the improvement of the accelerator luminosity, a quick feed-back of the information obtained by the event reconstruction to the accelerator operation is important, such as the precise position of the interaction point and the beam energy estimated from the event shape parameters. A detailed monitoring of the data quality is also important to keep high DAQ efficiency. Before RFARM was introduced, the monitoring was done using sampled events with limited software. RFARM enables the monitoring using the offline level reconstruction with a full data sample.

The monitoring is done based on the real time collection of histograms as shown in Fig. 6. The histograms are placed on a shared memory in each event processing node and the contents are periodically sent to the collection server through a socket connection. The contents are added on the server and then sent to the shared memory on the monitor server placed outside of the RFARM control network. The histograms are treated in the framework of the Belle Data Quality Monitor system (DQM) together with other histograms from event building farm. Fig. 7 shows some examples of the monitoring histograms obtained by RFARM.

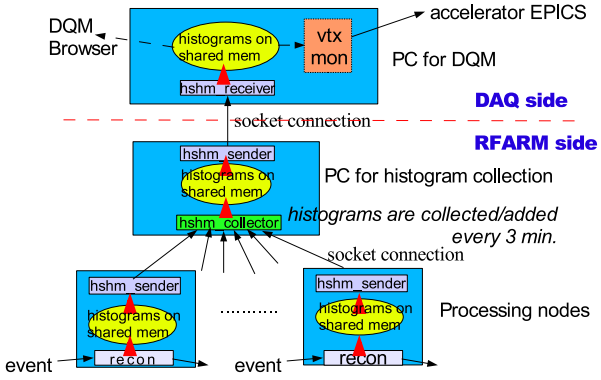

Figure 6: A schematic view of the histogram collection mechanism in RFARM.

SUMMARY

- The Real Time Event Reconstruction Farm (RFARM) is developed and implemented in the Belle DAQ system.
- RFARM is now being operated in the beam runs and performing the real time event reconstruction smoothly with the current DAQ condition. The offline-level event selection is not yet turned on.
- The data quality monitor using RFARM is now a part of the real time data monitoring system of Belle.

To cope with the further luminosity increase, the upgrade of the Belle data acquisition system is being planned as shown in Fig. 8. In the design, an event building farm and an RFARM are treated as a unit and the data are distributed

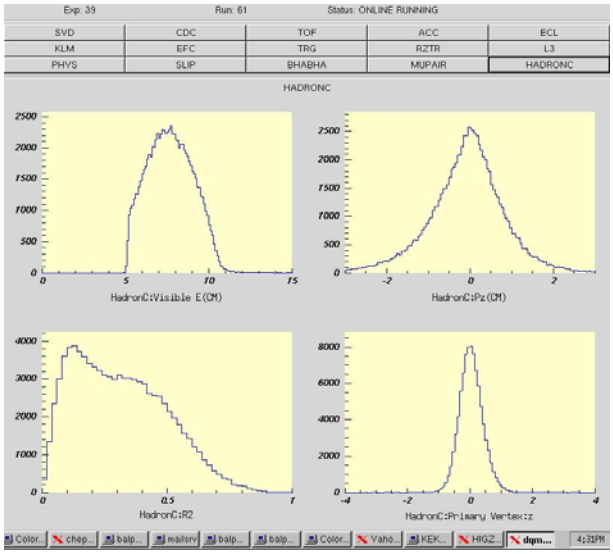

Figure 7: Histograms for monitoring obtained by RFARM (Various distributions for the hadronic events).

and processed by multiple units through the network matrix. The R&D on this new scheme is now in progress.

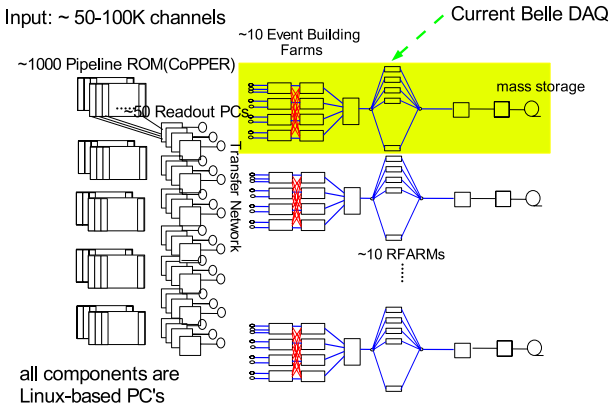

Figure 8: A design of the upgrade DAQ system.

REFERENCES

[1] M.Nakao, *et al.*, "Switchless Event Building Farm for Belle", IEEE Trans. on Nucl. Sci. **48**, 2385 (2001)

[2] R.Itoh, "BASF - Belle AnalysiS Framework", Proceedings of CHEP97, A244 (1997).

[3] M.Nakao and S.Y.Suzuki, " Network Shared Memory Framework for the Belle Data Acquisition Control", IEEE Trans. on Nucl. Sci. **47**, 267 (1999).